WHAT IS CLAIMED IS:

1. A method for processing audio data, comprising:

applying a plurality of anchor models to the audio data;

mapping the output of the plurality of anchor models into frame tags; and

producing the frame tags;

wherein the plurality of anchor models comprise a discriminatively-trained classifier.

2. The method as set forth in claim 1, wherein the discriminatively-trained classifier comprises a convolutional neural network classifier.

3. The method as set forth in claim 2, further comprising training the convolutional classifier on data separate from audio data available in a use phase.

4. The method as set forth in claim 2, wherein the convolutional neural network classifier is a time-delay neural network (TDNN) classifier.

5. The method as set forth in claim 4, further comprising training the TDNN classifier using cross entropy.

6. The method as set forth in claim 1, further comprising pre-processing the audio data to generate input feature vectors for the discriminatively-trained classifier.

7. The method as set forth in claim 1, further comprising normalizing a feature vector output of the discriminatively-trained classifier.

8. The method as set forth in claim 7, wherein the normalized feature vectors are vectors of unit length.

9. The method as set forth in claim 1, further comprising:

accepting a plurality of input feature vectors corresponding to audio features contained in the audio data; and

applying the discriminatively-trained classifier to the plurality of input feature vectors to produce a plurality of anchor model outputs.

10. The method as set forth in claim 1, wherein the mapping comprises:

clustering anchor model outputs from the discriminatively-trained classifier into separate clusters using a clustering technique; and

associating a frame tag to each separate cluster.

11. The method as set forth in claim 10, further comprising applying temporal sequential smoothing to the frame tag using temporal information associated with the anchor model outputs.

12. The method as set forth in claim 1, further comprising:

training the discriminatively-trained classifier using a speaker training set containing a plurality of known speakers; and

pre-processing the speaker training set and the audio data in the same manner to provide a consistent input to the discriminatively-trained classifier.

13. A computer-readable medium having computer-executable instructions for performing the method recited in claim 1.

14. A computer-implemented process for processing audio data, comprising:

applying a plurality of anchor models to the audio data;

mapping the output of the anchor models into frame tags; and

producing the frame tags;

wherein the plurality of anchor models comprise a discriminatively-trained classifier that is previously trained using a training technique.

15. The computer-implemented process of claim 14, wherein the training technique employs a cross-entropy cost function.

16. The computer-implemented process of claim 14, wherein the training technique employs a mean-square error metric.

17. The computer-implemented process of claim 14, wherein the discriminatively-trained classifier comprises a convolutional neural network classifier.

18. The computer-implemented process of claim 14, further comprising:

omitting an output non-linearity, which was used during training, from the discriminatively-trained classifier to generate a modified feature vector output; and

normalizing the modified feature vector output.

19. The computer-implemented process of claim 18, wherein normalizing further comprises creating a modified feature vector output having unit length.

20. A method for processing audio data containing a plurality of speakers, comprising:

applying a plurality of anchor models to the audio data;

mapping an output of the anchor models into frame tags; and

constructing a list of start and stop times for each of the plurality of speakers based on the frame tags;

wherein the plurality of anchor models comprise a discriminatively-trained classifier previously trained using a training set containing a set of training speakers, and wherein the plurality of speakers is not in the set of training speakers.

21. The method as set forth in claim 20, wherein the discriminatively trained classifier is a time-delay neural network (TDNN) classifier.

22. The method as set forth in claim 20, further comprising normalizing a feature vector output from the convolutional neural network classifier by mapping each element of the feature vector output to a unit sphere such that the feature vector output has unit length.

23. One or more computer-readable media having computer-readable instructions thereon which, when executed by one or more processors, cause the one or more processors to implement the method of claim 20.

24. A computer-readable medium having computer-executable instructions for processing audio data, comprising:

training a discriminatively-trained classifier in a discriminative manner during a training phase to generate parameters that can be used at a later time by the discriminatively-trained classifier;

applying the discriminatively-trained classifier that uses the parameters to the audio data to generate anchor model outputs; and

clustering the anchor model outputs into frame tags of speakers that are contained in the audio data.

25. The computer-readable medium of claim 24, further comprising pre-processing a speaker training set during the training and validation phase to

produce a first set of input feature vectors for the discriminatively-trained classifier.

26. The computer-readable medium of claim 25, further comprising pre-processing the audio data during the use phase to produce a second set of input feature vectors for the discriminatively-trained classifier, the pre-processing of the audio data being preformed in the same manner as the pre-processing of the speaker training set.

27. The computer-readable medium of claim 24, further comprising normalizing the feature vector outputs to produce feature vectors having a unit length.

28. The computer-readable medium of claim 27, wherein normalizing further comprises omitting a nonlinearity from the discriminatively-trained classifier during the use phase.

29. The computer-readable medium of claim 25, further comprising applying temporal sequential smoothing to the clustering the clustered feature vector outputs to produce the frame tags.

30. A computer-implemented audio processing process for segmenting and classifying speakers within audio data, comprising:

dividing the audio data into a plurality of frames;

using spectral analysis to extract feature vectors from each of the plurality of frames;

outputting a set of input feature vectors containing the extracted feature vectors;

applying a time-delay neural network (TDNN) classifier to the input feature vectors to produce anchor model outputs, the TDNN classifier having

been trained previously using a training technique and a speaker training set containing known speakers; and

generating frame tags from the anchor model outputs, such that each frame tag corresponds to a single one of the speakers.

31. The computer-implemented audio processing process as set forth in claim 30, further comprising extracting magnitudes from the feature vectors.

32. The computer-implemented audio processing process as set forth in claim 30, further comprising using Mel warped triangular filtering to average feature vector magnitudes and produce averaged feature vectors.

33. The computer-implemented audio processing process as set forth in claim 32, further comprising applying automatic gain control to the averaged feature vectors such that energy in each frame is smoothly kept at approximately a constant level.

34. The computer-implemented audio processing process as set forth in claim 32, further comprising applying dynamic thresholding to the averaged feature vectors to generate upper and lower energy boundaries.

35. The computer-implemented audio processing process as set forth in claim 30, wherein each of the plurality of frames has a duration of greater than 20 milliseconds.

36. A method for segmenting and classifying a plurality of speakers contained within audio data, comprising:

pre-processing the audio data by dividing the audio data into a plurality of frames having a duration of approximately 32 milliseconds;

using spectral analysis to extract feature vectors from each of the plurality of frames;

applying a time-delay neural network (TDNN) classifier to the feature vectors to produce anchor model outputs; and

extracting numbers from the TDNN classifier before the nonlinearity used in the training phase, to produce modified anchor model output vectors from the anchor model outputs.

37. The method of claim 36, further comprising adjusting the modified anchor model output vector such that each modified anchor model output vector has unit length in multi-dimensional space.

38. The method of claim 36, further comprises mapping the elements of the modified anchor model output vector to a unit sphere.

39. A method for processing audio data containing unknown speakers to segment and classify the speakers into separate and distinct classes, comprising:

pre-processing the audio data by dividing the audio data into a plurality of frames;

extracting feature vectors from each of the plurality of frames using spectral analysis;

applying a time-delay neural network (TDNN) classifier to the extracted feature vectors to produce anchor model outputs, the TDNN classifier having been previously trained using a speaker training set containing known speakers;

normalizing the anchor model outputs to produce unit length anchor model output vectors;

clustering unit length anchor model output vectors into initial frame tags;

selecting a frame tag for a point in time;

assigning a weight to neighboring frame tags of the selected frame tag; and

determining whether to change the initial classification of the selected frame tag based on the weights and neighboring frame tags.

40. An audio processing system, comprising:

audio data that contains a plurality of unknown speakers;

anchor models that comprise a discriminatively-trained classifier that inputs the audio data and produces anchor model outputs; and

a mapping module that maps the anchor model outputs to frame tags, such that each of the frame tags correspond to a single one of the plurality of unknown speakers in the audio data

41. The audio processing system as set forth in claim 40, wherein the discriminatively-trained classifier is a convolutional neural network classifier.

42. The audio processing system as set forth in claim 41, wherein the convolutional neural network is a time-delay neural network (TDNN) classifier.

43. The audio processing system as set forth in claim 40, further comprising an audio pre-processing module for pre-processing the audio data.

44. The audio processing system as set forth in claim 43, wherein the audio pre-processing module further comprises a frame module that divides the audio data into a plurality of input frames.

45. The audio processing system as set forth in claim 44, wherein each of the plurality of input frames has a duration of at least 32 milliseconds.

46. The audio processing system as set forth in claim 44, wherein the audio pre-processing module further comprises a spectral features processor that extracts spectral feature magnitudes from each of the plurality of input frames.

47. The audio processing system as set forth in claim 40, further comprising a normalization module that produces normalized anchor model output vectors.

48. The audio processing system as set forth in claim 47, wherein the normalization module further comprises a unit sphere module that maps the modified anchor model output vectors to a unit sphere to produce anchor model output vectors having unit length.

49. The audio processing system as set forth in claim 40, wherein the mapping module further comprises a temporal sequential smoothing module for reducing frame tag error.

50. The audio processing system as set forth in claim 49, wherein the temporal sequential smoothing module further comprises a data selection module that inputs a set of frame tags produced by the mapping module and selects a frame tag from the set.

51. The audio processing system as set forth in claim 50, wherein the temporal sequential smoothing module further comprises a neighbor examination module that assign weights to each of the neighboring frame tags of the selected frame tag.

52. The audio processing system as set forth in claim 51, wherein the temporal sequential smoothing module further comprises a frame tag correction module that changes the selected frame tag based on weights and neighboring frame tags.

53. An audio processing system for classifying speakers within audio data, comprising:

a speaker training set containing known speakers;

a training system, producing parameters for a time-delay neural network (TDNN) classifier that distinguishes between the known speakers;

audio data containing unknown speakers;

a speaker classification system, comprising:

an audio pre-processing module for dividing the audio data into a plurality of frames and producing input feature vectors;

a TDNN classifier that uses the parameters produced by the training system and applies them to the input feature vectors to produce anchor model output vectors; and

a clustering module that clusters anchor model output vectors and assigns frame tags to each anchor model output vector.

54. The audio processing system as set forth in claim 53, wherein the training system comprises a cross-entropy error module.

55. The audio processing system as set forth in claim 54, wherein the training system comprises a mean-squared error module.

56. The audio processing system as set forth in claim 53, wherein each of the plurality of frames has a duration of at least 32 milliseconds.

57. The audio processing system as set forth in claim 53, wherein the clustering module further comprises a temporal sequential smoothing module that selects a frame tag and determines whether to change the frame tag based on weights and neighboring frame tags.

58. The audio processing system as set forth in claim 53, wherein each of the frame tags is associated with an input frame for the TDNN classifier and wherein each input frame contains approximately 1 second of the audio data.

59. The audio processing system as set forth in claim 53, wherein the audio pre-processing module further comprises an input frame of approximately 1 second duration that the TDNN classifier sees as input to the TDNN classifier.

60. The audio processing system as set forth in claim 59, wherein the input frame is associated with a frame tag.